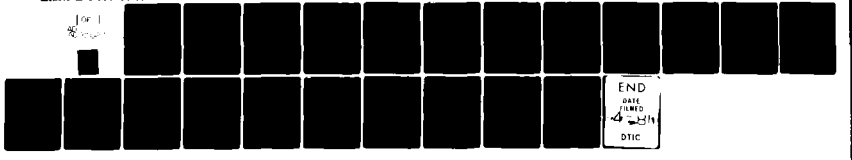
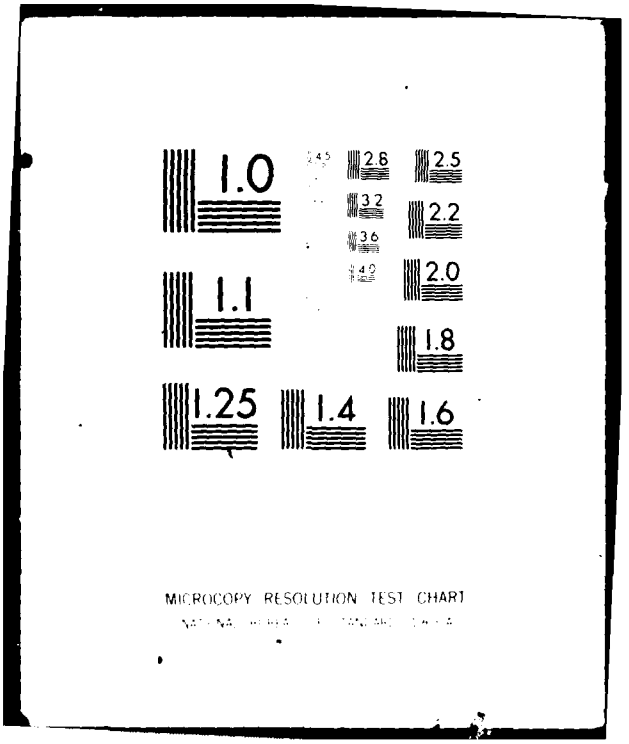
AD-A097 027 WISCONSIN UNIV-MADISON DEPT OF STATISTICS F/G 12/1
CROSS VALIDATION AND CONSTRAINED REGULARIZATION METHODS FOR MIL--ETC(U)
JAN 80 G WAHBA N00014-77-C-0675
UNCLASSIFIED UWIS-DS-80-629 NL




1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6
MICROCOPY RESOLUTION TEST CHART

DEPARTMENT OF STATISTICS
University of Wisconsin
1210 W. Dayton St.
Madison, WI 53706

LEVEL II

(12)

TECHNICAL REPORT NO. 629

January 1980

CROSS VALIDATION AND CONSTRAINED REGULARIZATION METHODS FOR MILDLY ILL POSED PROBLEMS

by

Grace Wahba

University of Wisconsin, Madison

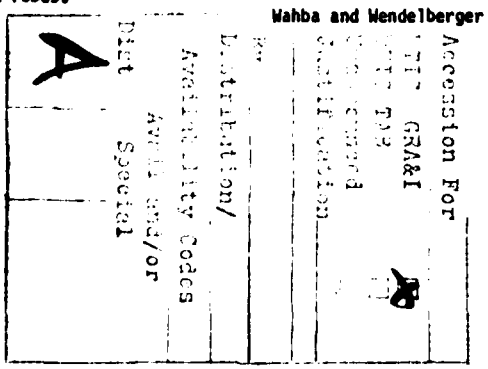
This research supported by the Office of Naval Research under Contract No. N00014-77-C-0675.

Prepared for the Proceedings of the International Conference on Ill Posed Problems, held at Newark, Delaware, October 2-6, 1979. M.Z. Nashed, Ed.

This paper is a shortened and revised version of Department of Statistics, University of Wisconsin-Madison Technical Report No. 595.

ABSTRACT

The model $z_i = \int_0^1 K(t_i,s)f(s)ds + \varepsilon_i$, $i = 1,2,\ldots,n$, is studied, where K is assumed known, the $\{\varepsilon_i\}$ are random errors, and it is desired to estimate f from the data (z_i). The (regularized) estimate of f will be taken as the minimizer in H of $\frac{1}{n}\sum_{i=1}^{n}((Kf)(t_i)-z_i)^2 + \lambda||f||^2$ (=(*)), where H is a reproducing kernel space with norm $||\cdot||$. We first define the intrinsic rank of this experiment. This definition is used to provide insight into the circumstances in which one may expect to estimate f well, moderately well, or poorly. The sensitivity of a regularized estimate to the $\{\varepsilon_i\}$ is made explicit. After giving the intrinsic rank of the examples of first and second derivative, Abel's equation and Fujita's equation, it is argued that the first three are only mildly ill posed and if f is "nice" it will be amenable to accurate estimation. The method of generalized cross validation for choosing λ is reviewed and numerical results for the estimation of first and second derivative from noisy data are given. A method for solving Abel's equations when inversion formulae are available is described which exploits properties of the above mentioned first derivative estimate. The use of outside information in the estimation of f is then considered. We consider the cases where f is known to satisfy a finite number of linear equality constraints (for example, boundary values, or moment conditions) or a continuous family of linear inequality constraints (including positivity, monotonicity or convexity). f is then estimated by minimizing or approximately minimizing (*), subject to these constraints. We extend the generalized cross validation method for choosing λ to this case. Finally, the extension of this method to certain robust estimation problems is suggested.



1. Introduction

In this paper we consider the model

$$z_i = \int_0^1 K(t_i,s)f(s)ds + \varepsilon_i, \quad i = 1,2,\ldots,n, \tag{1.1}$$

where $z = (z_1,\ldots,z_n)'$ is the data vector, $K(t,s)$ is known exactly, $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_n)'$ is a vector of independent zero mean random variables with common (unknown) variance σ^2, and it is desired to estimate f, given the observations z. We consider the estimation of f by $f_{n,\lambda}$, the solution to the problem: Find f in H to minimize

$$\frac{1}{n}\sum_{i=1}^{n}((Kf)(t_i)-z_i)^2 + \lambda||f||^2, \tag{1.2}$$

where H with the norm $||\cdot||$ is a reproducing kernel Hilbert space and the map $f\to(Kf)(t)$ is a bounded linear functional on H for each fixed $t = t_1,\ldots,t_n$. Later we will consider minimizing (1.2) subject to $f\in C$, where C is a closed, convex set in H, which may, with appropriate H, be chosen to reflect outside information such as: f satisfies given boundary or moment conditions (exactly), or f is non-negative, or monotone, or convex. We will show how the method of generalized cross validation (GCV) can be used to estimate the regularization parameter in these cases.

The minimization of (1.2), with H suitably chosen ($H = W_2^m$, m = 1 or 2 for example), and λ chosen by the GCV method, has given satisfactory results in a number of studies (see, for example, Bjorck and Elden (1979) Colli Franzone et al (1979), Merz (1978, 1979a), Utreras (1979a,b) Wahba and Wendelberger (1980). (m can also be chosen by GCV, see Wahba and Wendelberger (1980)). However, if one expects to estimate f from

z alone, then the data z must contain sufficient information concerning f to meet the experimenter's requirements, before any mathematical technique can be expected to "wring it out". If the data contains this information, then satisfactory solutions can be obtained, usually with the aid of sophisticated mathematical techniques and a powerful computer, and we shall call such problems mildly ill posed. If not, then sophisticated techniques and powerful computers will not provide the missing information (contrary to uninformed belief!). If the low information content of the experiment can be recognized, then at least two routes are available - a) a redesign of the experiment to provide more information, b) the incorporation of a priori or outside information into the solution.

In Section 2 we define the "intrinsic rank" of the experiment described by (1.1). The intrinsic rank of an experiment is the number of linearly independent pieces of information practically available in the data vector z about the function f if there were no errors in the data and K is known perfectly. Examination of the intrinsic rank of a problem can provide valuable information concerning whether or not a satisfactory solution is obtainable. It is computable for the problems we consider and should be done routinely.

In Section 3 we discuss the effect of noise on the solution. This is most easily done in terms of what we shall call the canonical representers. The estimated solution will always be in the span of the canonical representers, and so knowledge of them can be a useful diagnostic tool if problems appear. We note here that the intrinsic rank as well as the canonical representers depend on H as well as n, K, and the location of the t_i's.

In Section 4 we give the intrinsic rank, as a function of n, for the examples of first derivative, k^{th} derivative, Abel's equation and Fujita's equation. It can be seen that first and second derivatives, and solutions of Abel's equation should be usefully recoverable with reasonable data sets, while estimation of f by solving Fujita's equation is hopeless in the geometry that we considered.

In Section 5, we first review the method of generalized cross validation for choosing λ, and some of its properties. We describe some successful experiments in which first and second derivatives were well estimated from noisy data, using smoothing splines with GCV for choosing the smoothing parameter. We then propose a new class of methods for solving Abel-type equations given noisy data, when inversion formulae are available. These methods exploit the ability of smoothing splines with GCV to estimate derivatives well.

In Section 6 we extend the idea of cross validation to problems where f is known to be in a closed convex set C. The case when C consists of functions {f} which satisfy a finite number of linearly equality constraints, (for example boundary or moment conditions) is discussed in Section 6.2. The extention of the GCV method is straightforward, and it seems fairly evident that the properties of the method can be established rigorously in this case. However, in this paper only suggestive arguments, and not proofs, are given. In Section 6.3 we consider C described by a family of linear inequality constraints satisfying certain continuity conditions. This includes the cases where

C consists of nonnegative, or convex, or monotone functions. When the family of linear inequality constraints determining C can be approximated by a finite family of inequality constraints (for example, $f(\frac{i}{N}) \geq 0$, $i = 0,1,\ldots,N$) then the (approximate) constrained solution is also the minimizer of (1.2) subject to a finite number of linear equality constraints (namely, the active constraints in the inequality constrained solution). In this case, an approximate GCV estimate of λ can be computed in a straightforward manner.

This paper has been considerably shortened from the original conference paper, in particular, reference to $||\cdot||$ a seminorm instead of a norm has been omitted. Other topics omitted are a discussion of basis function methods, tests of the adequacy of the model (1.1), and applications to computerized tomography. The reader may consult Wahba (1979d) for details.

2. *The intrinsic rank of an ill posed experimental problem*

The intrinsic rank of an ill posed experimental problem as we define it here depends on the following:

i) the operator K

ii) the number and location of the data points $t_1,\ldots,t_n$

iii) the space H in which the solution is sought

iv) the minimum computer roundoff, δ(i.e. $\delta \sim 10^{-16}$ in double precision.)

The intrinsic rank r_I will be the useful number of linearly independent pieces of information about f in the absence of measurement errors, errors in K or cumulative roundoff beyond that in iv).

The effective rank will be less than r_I and will depend on the above as well as

i) σ^2

ii) errors in knowledge of K

iii) roundoff errors beyond iv) above.

Errors in knowledge of K can be an important source of trouble and it is important to consider this source of error when dealing with experimental data. However, in this work we assume K is known accurately. We will also assume that computer roundoff (iii) and any quadrature error will be made negligible compared to experimental error (σ^2) by careful tailoring of the numerical methods used to the intrinsic rank of the problem. The effect of σ^2 is discussed in Section 3. We shall ignore roundoff in the discussion. (We believe that what is sometimes taken for numerical error is actually the division of a random error by a very small eigenvalue, see Section 3).

We now prepare to define the intrinsic rank r_I of the problem (1.1). We suppose that f is estimated by $f_{n,\lambda}$, the solution to the problem: Find $f \in H$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(z_i - \int_0^1 K(t_i,s)f(s)ds)^2 + \lambda||f||^2, \qquad (2.1)$$

where $||f||$ is the norm of f in H. The popular case where $||\cdot||$ is a seminorm, for example $||f|| = [\int_0^1 (f''(t))^2 dt]^{1/2}$, can be treated similarly but is omitted for lack of space. Some details may be found in Wahba (1979d). It is required that H be a space in which the n functionals which map f to $\int_0^1 K(t_i,s)f(s)ds$, $i = 1,2,\ldots,n$, are continuous linear functions. If this is the case, then by the Riesz representation theorem (Akhiezer and Glazman

(1961)), there exist n elements $\eta_1,\ldots,\eta_n$, in H called the representers, such that

$$\int_0^1 K(t_i,s)f(s)ds = \langle \eta_i, f\rangle,\ f\in H,\ i = 1,2,\ldots,n$$

where $\langle\cdot,\cdot\rangle$ is the inner product in H. For example if $H = L_2[0,1]$ then for fixed i,

$$\eta_i(s) = K(t_i,s)\quad i = 1,2,\ldots,n.$$

If $H = H_R$, the r.k.h.s. with r.k. R(s,t), then

$$\eta_i(s) = \int_0^1 K(t_i,u)R(s,u)du. \tag{2.2}$$

The reproducing kernel space results we use in this paper can be found in Kimeldorf and Wahba (1971), (where the seminorm case is considered in detail), see also Aronszajn (1950). If H is a finite dimensional space, then each η_i is a linear combination of basis functions.

The solution $f_{n,\lambda}$ to the minimization problem of (2.1) can be written

$$f_{n,\lambda} = \tilde{K}_n^*(\tilde{K}_n\tilde{K}_n^* + n\lambda I)^{-1}z,$$

where $z = (z_1,\ldots,z_n)'$. $\tilde{K}_n$ is the operator which maps H into E_n as follows:

$$\tilde{K}_n f = \begin{pmatrix} \int_0^1 K(t_1,s)f(s)ds \\ \vdots \\ \int_0^1 K(t_n,s)f(s)ds \end{pmatrix}.$$

$\tilde{K}_n^*$ is the adjoint of $\tilde{K}_n$ in the sense that $\tilde{K}_n^*\colon E_n \to H$, and $\tilde{K}_n^*$ satisfies

$$(z,\tilde{K}_n f) = \langle \tilde{K}_n^* z, f\rangle\ ,$$

where $(\cdot,\cdot)$ is the Euclidean inner product. It can be verified that $\tilde{K}_n^*z$ has the representation

$$(\tilde{K}_n^*z)(s) = \sum_{i=1}^n z_i\eta_i(s).$$

$\tilde{K}_n\tilde{K}_n^*\colon E_n \to E_n$ is the operator of multiplication by the $n\times n$ matrix with jk^{th} entry $\langle \eta_j,\eta_k\rangle$. This matrix is the Gram matrix of the representers of the data functionals. If $H = L_2[0,1]$, then

$$\langle \eta_j,\eta_k\rangle = \int_0^1 \eta_j(s)\eta_k(s)ds = \int_0^1 K(t_j,s)K(t_k,s)ds,$$

and, if $H = H_R$, then

$$\langle \eta_j,\eta_k\rangle = \int_0^1\int_0^1 K(t_j,s)R(s,t)K(t_k,t)dsdt. \tag{2.3}$$

The matrix $(\tilde{K}_n\tilde{K}_n^*)$ is symmetric non-negative definite, and hence has a decomposition

$$(\tilde{K}_n\tilde{K}_n^*) = \Gamma D\Gamma' \tag{2.4}$$

where Γ is an $n\times n$ orthogonal matrix and D is a diagonal matrix with eigenvalues (diagonal entries) $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$.

We define the intrinsic rank r_I as the number of eigenvalues λ_i for which $\lambda_i/\lambda_1 > 6\times 10^{-16}$ (computer roundoff). Thus r_I is the effective number of linearly independent data functionals in the experiment (1.1) in the absence of experimental errors or errors in K.

We make several observations about r_I. Firstly, *if H is a finite* dimensional space of dimension N, then $r_I \leq N$. This is reasonable, since, if f is known to be in H, then f is determined by N linearly independent pieces of information, and the experiment (1.1) cannot deliver more. Secondly, if H is a space of functions with several continuous derivatives, then (other things being equal), r_I will be less than if H is L_2. Again, this is reasonable, since, loosely speaking, $\int K(t_i,s)f(s)ds$ and $\int K(t_j,s)f(s)ds$ can be expected to be less linearly independent on smooth functions than on arbitrary elements of L_2.

3. The effect of noise. The canonical representers

Let Γ and $\lambda_1,\ldots,\lambda_n$ be defined by (2.4). We define the canonical data vector $y = (y_1,\ldots,y_n)'$ and the canonical representers $\phi_1,\ldots,\phi_n$ by

$$y = \Gamma' z$$

$$\begin{pmatrix} \phi_1(s) \\ \vdots \\ \phi_n(s) \end{pmatrix} = \Gamma' \begin{pmatrix} \eta_1(s) \\ \vdots \\ \eta_n(s) \end{pmatrix}.$$

Then

$$\langle \phi_j, \phi_k \rangle = \lambda_j, \ j = k$$

$$= 0 \quad j \neq k.$$

Since

$$z_j = \langle \eta_j, f \rangle + \epsilon_j, \quad j = 1,2,\ldots,n$$

we have

$$y_j = \langle \phi_j, f \rangle + \tilde{\epsilon}_j$$

where

$$\begin{pmatrix} \tilde{\epsilon}_1 \\ \vdots \\ \tilde{\epsilon}_n \end{pmatrix} = \Gamma' \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

and, if the ϵ_i are normally distributed, the $\tilde{\epsilon}_i$ are independent, normally distributed with mean 0 and variance σ^2. Since the experiment (1.1) provides the equivalent data y, one "knows" the inner product of f with the unit vector $\psi_i = \phi_i/\sqrt{\lambda_i}$ from the data to an accuracy of, within, say ±2 standard deviations $= 2\sigma/\sqrt{\lambda_i}$. That part of f not in the span of the first r_I canonical representers is not "seen" by the experiment, even in the absence of measurement error.

4. The intrinsic rank of some examples

In these examples we suppose the t_i are equally spaced, although no doubt the results are true for $\max_i|t_{i+1}-t_i|/\min_i|t_{i+1}-t_i|$ bounded.

4.1 The first derivative, $H = H^1$

Let $H = H^1$, $H^1 = \{f: f \text{ abs. cont.}, f' \in L_2[0,1]\}$, and let

$$(Kf)(t) = \int_0^t f(s)ds,$$

thus, $\frac{d}{dt}(Kf)(t) = f(t)$. Here $\lambda_n/\lambda_1 = O(n^{-4})$. The rate $O(n^{-4})$ is obtained as follows: The reproducing kernel for H^1 with $\|f\|^2 = f^2(0) + \int_0^1 (f'(u))^2 du$ is

$$R(s,t) = 1 + \min(s,t).$$

Define

$$P(s,t) = \int_0^1\int_0^1 K(s,u)R(u,v)K(t,v)dudv$$

$$= \int_0^s\int_0^t (1+\min(s,t))dsdt.$$

Then $\tilde{K}_n\tilde{K}_n^*$ is the operator of multiplication by the n×n matrix with jk^{th} entry $P(t_j,t_k)$. P is a Green's function for a 4th order linear differential operator, thus the eigenvalues of the Hilbert Schmidt operator with

kernel P, are $O(n^{-4})$, i.e. inversely related to the eigenvalues of the associated differential operator. An argument in Craven and Wahba (1979), see also Wahba (1977, 1979c) indicates that the eigenvalues of the matrix obtained by discretizing P behave roughly like n times the eigenvalues of P, giving $\lambda_n/\lambda_1 = O(n^{-4})$. For a carefully developed argument which gives similar results in a related problem, see Utreras (1979b).

If $n = 10^3$ then $\lambda_n/\lambda_1 = 10^{-12}$, and for $\delta = 10^{-14}$, say, $\tilde{K}_n\tilde{K}_n^*$ is of full intrinsic rank. Provided that the data are not too noisy, this indicates that effective numerical differentiation is feasible if the true f is reasonably "nice".

4.2 The k^{th} derivative, $H = H^m$

Let $H = H^m$, $H^m = \{f: f, f', \ldots, f^{(m-1)}$ abs. cont., $f^{(m)} \in L_2[0,1]\}$ and let

$$(Kf)(t) = \int_0^t \frac{(t-s)^{k-1}}{(k-1)!} f(s)ds,$$

thus $\frac{\partial^k}{\partial t^k}(Kf)(t) = f(t)$. Here $\lambda_n/\lambda_1 = O(n^{-2(m+k)})$. (Reproducing kernels for H^m may be found in, for example, Kimeldorf and Wahba (1971).

For example, if $k = 3$, $m = 2$, then $\lambda_n/\lambda_1 = O(n^{-10})$. If $n \sim 10^{1.4} = 25$, then $\lambda_n/\lambda_1 \sim 10^{-14}$ and so the intrinsic rank of this problem will be around 25. If f is a very smooth function without much structure one might expect to get a "good picture" of f with 25 pieces of information. More precisely, if f is in the span of the first 25 canonical representers and σ^2 is not too big, then a useful estimate of f might be recoverable. Otherwise it probably won't be.

This indicates, however, that accurate estimation of second derivative (k=2), with m=1 is feasible with good quality data, since in this case $\lambda_n/\lambda_1 = O(n^{-4})$ and $\tilde{R}_n\tilde{R}_n^*$ will be of full effective rank for n as large as 150.

4.3 Abel's equations

These equations are of the form

$$(Kf)(t) = \int_t^b \frac{k(t,s)}{(s-t)^\alpha} f(s)ds$$

where k(t,s) is continuous on $0 \leq s \leq 1$ and $0 < \alpha < 1$. They behave like the equations in Section 4.2 with $k = 1 - \alpha$, and so they are of higher intrinsic rank than comparable problems involving numerical differentiation. If $H = H^m$, then $\lambda_n/\lambda_1 = O(n^{-2(m+1-\alpha)})$.

4.4 Fujita's equation and other severely ill posed problems

Fujita's equation relates centrifuge data to particle mass distribution. See Gehatia and Wiff (1970).

$$(Kf)(t) = \int_0^{s_{max}} \frac{\theta e^{-\theta s t}}{1-e^{-\theta s}} f(s)ds \quad t \in [0, t_{max}].$$

With $\theta = 4.25$ and realistic values of s_{max} and t_{max} we found this innocuous looking equation to be severely ill posed. With n = 41 equally spaced data points, and $H = H^1$, we computed the eigenvalues $\lambda_1, \ldots, \lambda_{41}$. They turned out to look roughly as in the following table

ν	λ_ν	ν	λ_ν
1	1	6-24	10^{-14} to 10^{-15}
2	$10^{-3.5}$	25-41	-10^{-15} to -10^{-14}
3	10^{-7}		
4	$10^{-10.5}$		
5	10^{-14}		

We concluded that eigenvalues 5 or 6 through 41 were "machine 0". The intrinsic rank of this problem is between 4 and 5.

The GCV estimate of the optimal λ (the GCV estimate is defined in the next section) was around 10^{-11} and was a very good estimate of the optimal λ as measured by how close it came to minimizing

$$T_D(\lambda) = \frac{1}{n}\sum_{i=1}^{41}(f_{n,\lambda}(\tfrac{i}{41}t_{max})-f(\tfrac{i}{41}t_{max}))^2$$

in an experiment with synthetic data where f was known. The estimate $f_{n,\lambda}$ may be written in terms of the canonical data and canonical representers of Section 3 as

$$f_{n,\lambda} = \sum_{i=1}^{n}\frac{y_i}{\lambda_i+\lambda}\phi_i .$$

Note that a λ of 10^{-11} is completely negligible compared to eigenvalues 1-3, and completely swamps eigenvalues 5-41. We succeeded in obtaining excellent solutions in some examples and nonsense results in others. See Wahba (1979c). We came to the conclusion that the excellent solutions occurred when f was effectively in the span of the first 4 canonical representers and the lousy results occurred when it was not.

Numerical inversion of the Laplace transform can be expected to be similarly nasty. The problem of inversion of radiance measurements (z) to obtain temperature profiles (f) from satellite radiance measurements in the NIMBUS 6 satellite and others also appears to be severely ill posed. See Smith and Wolfe (1976), Fritz et al (1972).

5. Cross validation methods for mildly ill posed problems

5.1 The method of generalized cross validation (GCV) for choosing λ

We review this method since it plays a role in the remainder of the paper. The theory has been developed in Wahba (1977), Craven and Wahba (1979) and Golub, Heath and Wahba (1979). We will refer to these last two

as CW and GHW respectively. Numerical results concerning the method are given or mentioned in CW, GHW, Utreras (1979a,b), Merz (1979a), Welch (1979), Bjorck and Elden (1979), Stutzle (1977), Colli Franzone et al (1979).

The idea is as follows: Let $f_{n,\lambda}^{[k]}$ be the minimizer of

$$\frac{1}{n}\sum_{\substack{i=1\\ i\neq k}}^{n}((Kf)(t_i)-z_i)^2 + \lambda||f||^2,$$

where $||\cdot||$ may be a norm or seminorm in H. If λ is a good choice, then $(Kf_{n,\lambda}^{[k]})(t_k)-z_k$, should, on average be small. This is measured by the ordinary cross validation function $V_0(\lambda)$ given by

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n}[(Kf_{n,\lambda}^{[k]})(t_k)-z_k]^2, \qquad (5.1.1)$$

and the ordinary cross-validation estimation of λ is the minimizer of $V_0(\lambda)$. It is shown in CW and GHW that

$$Kf_{n,\lambda}^{[k]}(t_k)-z_k \equiv [(Kf_{n,\lambda})(t_k)-z_k]/(1-a_{kk}(\lambda)) \qquad (5.1.2)$$

where

$$a_{kk}(\lambda) = \frac{\partial}{\partial z_k}(Kf_{n,\lambda})(t_k) \qquad (5.1.3)$$

and $f_{n,\lambda}$ is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}((Kf)(t_i)-z_i) + \lambda||f||^2. \qquad (5.1.4)$$

Hence

$$V_0(\lambda) \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{[(Kf_{n,\lambda})(t_k)-z_k]^2}{(1-a_{kk}(\lambda))^2} . \qquad (5.1.5)$$

It is also shown in CW and GHW that, from the point of view of minimizing predictive mean square error (defined later), $V_0(\lambda)$ should be replaced by

the generalized cross validation function (GCVF) $V(\lambda)$ defined by

$$V(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{((Kf_{n,\lambda})(t_k)-z_k)^2}{(1-a_{kk}(\lambda))^2}\omega_k(\lambda) \qquad (5.1.6)$$

where the weights $\omega_k(\lambda)$ are defined by

$$\omega_k(\lambda) = (1-a_{kk}(\lambda))/(1-\frac{1}{n}\sum_{j=1}^{n}a_{jj}(\lambda)). \qquad (5.1.7)$$

The GCV estimate λ is the minimizer of $V(\lambda)$. The weights $\omega_k(\lambda)$ compensate for the unequal influence of different data points; if the $a_{kk}(\lambda)$ are equal then $V(\lambda) = V_0(\lambda)$. For computational purposes we note that $a_{kk}(\lambda)$ is the kkth entry of the n×n matrix $A(\lambda)$ satisfying

$$\begin{pmatrix}(Kf_{n,\lambda})(t_1)\\ \vdots \\ (Kf_{n,\lambda})(t_n)\end{pmatrix} = A(\lambda)z. \qquad (5.1.8)$$

For computational purposes, then, substituting (5.1.7) and (5.1.8) into (5.1.6) results in

$$V(\lambda) = \frac{\frac{1}{n}||(I-A(\lambda))z||^2}{(\frac{1}{n}Tr(I-A(\lambda)))^2}. \qquad (5.1.9)$$

(Numerical methods for computing and minimizing $V(\lambda)$ are discussed in Bjorck and Elden (1979), Utreras (1979a,b), Wahba and Wendelberger (1980), Wendelberger (1980), see also Section 6 below). For purposes of later generalization we observe that $V(\lambda)$ of (5.1.2) may be rewritten as

$$V(\lambda) = \frac{\frac{1}{n}RSS(\lambda)}{(1-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial z_k}(Kf_{n,\lambda})(t_k))^2} \qquad (5.1.10)$$

where

$$RSS(\lambda) = \sum_{i=1}^{n}((Kf_{n,\lambda})(t_i)-z_i)^2.$$

The predictive mean square error $T(\lambda)$ is defined by

$$T(\lambda) = \frac{1}{n}\sum_{i=1}^{n}((Kf_{n,\lambda})(t_i)-(Kf)(t_i))^2$$

where f is the true answer in the model (1.1). Loosely speaking, for large n

$$EV(\lambda) \simeq ET(\lambda) + \sigma^2$$

for λ near the minimizer of $T(\lambda)$, see CW, CHW, Wahba (1977). (Here E is expectation.) There are other, possibly more desirable optimality criteria for λ, for example the minimization of

$$T_D(\lambda) = \int_0^1 (f_{n,\lambda}(t)-f(t))^2dt,$$

see also Nashed (1979b). One can obtain estimates for λ from the data which in theory (approximately) minimize T_D. Doing this in itself can, however, be an ill posed problem. In our numerical experiments with synthetic data we have generally found that the minimizers of $T_D(\lambda)$ and $T(\lambda)$ tend to be close, and $\hat{\lambda}$, the minimizer of $V(\lambda)$, is generally a good estimate of the minimizer of $T(\lambda)$. For this reason we have not seriously attempted to modify the optimality criteria. In a synthetic experiment, the inefficiency of $\hat{\lambda}$ can be measured by

$$T(\hat{\lambda})/\min_{\lambda}T(\lambda) \quad (\text{or } T_D(\hat{\lambda})/\min_{\lambda}T_D(\lambda)).$$

5.2 Examples

5.2.1 Estimation of the first derivative

Here the model is

$$z_i = g(t_i) + \epsilon_i,\ i = 1,2,\ldots,n \qquad (5.2.1)$$

where the ϵ_i are as before and $g \in H^2$. It is desired to estimate g'. We let $g_{n,\lambda}$ be the minimizer in H^2 of

$$\frac{1}{n}\sum_{i=1}^{n}(g(t_i)-z_i)^2 + \lambda\int_0^1 (g''(u))^2 du \qquad (5.2.2)$$

and estimate g' by $g_{n,\hat{\lambda}}'$, where $\hat{\lambda}$ is the minimizer of $V(\lambda)$ of (5.1.2). $g_{n,\lambda}$ is the cubic polynomial smoothing spline discussed in Reinsch (1967) and is differentiated analytically. Successful numerical results appear in CW, Merz (1978), Utreras (1979a,b), and elsewhere.

Transportable code is available from Merz (1978), Utreras (1979b) and Fleisher (1979). Our experience with the method indicates that it will do well for $n \geq 20$ or so, whenever g is "smooth", there are at least 7 or 8 data points per local maximum in g' and when σ is of the order of a fraction of a percent to several percent of the range of g. Code for computing $g_{n,\hat{\lambda}}$ is available in IMSL (1980), version 8 subroutine ICSSCV.

5.2.2 Estimation of the second derivative. Numerical results

If g in (5.2.1) has a smooth second derivative, it can be estimated by differentiating $g_{n,\hat{\lambda}}$ twice. This should give good results in the interior of $[0,1]$, however $g_{n,\lambda}''(0) = g_{n,\lambda}''(1) = 0$, for any λ, so that one cannot estimate $g''(t)$ for t in a small neighborhood of 0 or 1 unless $g''(0) = g''(1) = 0$. This problem at the boundary can be eliminated by using quintic splines, that is, by replacing $\int_0^1 (g''(u))^2 du$ by $\int_0^1 (g''(u))^2 du$ in (5.2.2), resulting in a quintic smoothing spline. A fast algorithm for the quintic smoothing spline with the GCV estimate of λ has been developed by Utreras (1979a).

A Monte Carlo example of the estimation of second derivative of a periodic function in the presence of noisy data appears in Wahba (1979c), and we reproduce the example. The results were fairly typical of a large

number of similar unpublished examples with high quality (Monte Carlo) data. In this example

$$g(t) = \int_0^1 K(t,s)f(s)ds$$

with

$$K(t,s) = \frac{1}{2}\{|t-s|^2-|t-s|+\frac{1}{6}\}.$$

$K(t,s)$ is a Green's function for the second derivative operator such that, if $g = Kf$, then g is the solution to $g'' = f$, $\int_0^1 g(y)dy = 0$, $g(0) = g(1) = 0$. The solid line in Figure 1a is g and the cross marks are the data $z_i = g(t_i) + \epsilon_i$ where the ϵ_i were simulated normally distributed errors, with variance σ^2. σ was about 1/300 of the range of g. f is estimated as $f_{n,\lambda}$, the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}((Kf)(t_i)-z_i)^2 + \lambda\int_0^1 (f'(u))^2 du$$

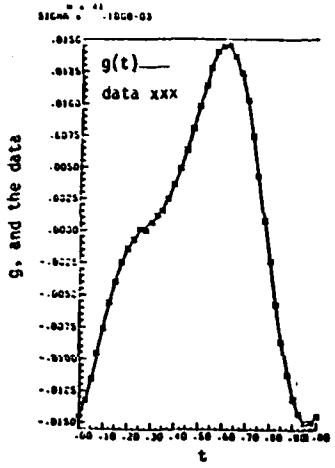
in the subspace of H^1 satisfying the (periodic) conditions $\int_0^1 f(u)du = 0$, $f(0) = f(1)$. The true f also satisfied these conditions. λ was chosen to minimize $V(\lambda)$. The calculation is that suggested in Wahba (1977), where the fact that $\|\cdot\|$ is a norm on the periodic functions considerably simplified the expressions. $V(\lambda)$ is plotted in 1b along with the mean square errors $T(\lambda)$ and $T_D(\lambda)$ defined by

$$T(\lambda) = \frac{1}{n}\sum_{i=1}^{n}((Kf_{n,\lambda})(t_i) - (Kf)(t_i))^2$$

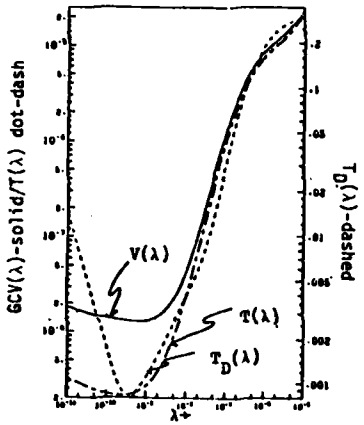
and

$$T_D(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(f_{n,\lambda}(t_i)-f(t_i))^2.$$

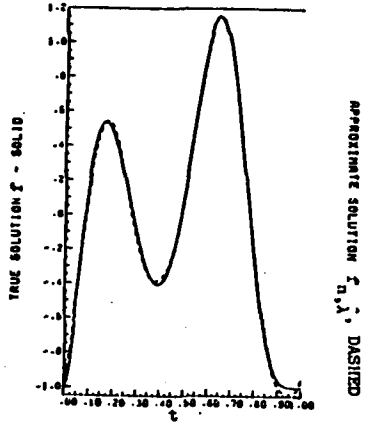
It can be seen that the minimizer of $V(\lambda)$ is a good estimate of the minimizer of both $T(\lambda)$ and $T_D(\lambda)$. The theory in CW and GWH says that $V(\lambda)$ should "track" $T(\lambda)$ and one can see that this does in fact happen. Figure 1c compares the true and estimated second derivative. It can be seen that the results are very good.



(a) The true g and the data



(b) The GCV function $V(\lambda)$ and the mean square errors $T(\lambda)$ and $T_D(\lambda)$ in $Kf_{n,\lambda}$ and $f_{n,\lambda}$.



(c) The true and estimated second derivative.

Figure 1. Estimation of Second Derivative

5.2.3 Abel's equations

These equations have been studied by Anderssen (1976), Anderssen and de Hoog (1979), Anderssen and Jakeman (1975) and Jakeman and Anderssen (1975a,b). They have provided solution methods and a number of numerical results. Anderssen and de Hoog (1979) have called these problems "weakly ill posed". Some of these equations have inversion formulae involving the first derivative. For example (Anderssen (1976)) if

$$g(t) = 2 \int_t^{t_{max}} \frac{sf(s)}{(s^2-t^2)^{1/2}}\, ds \qquad (5.2.3)$$

then

$$f(s) = -\frac{1}{\pi} \int_s^{t_{max}} \frac{g'(t)}{(t^2-s^2)^{1/2}}\, dt. \qquad (5.2.4)$$

In addition to the spectral differentiation - product integration methods proposed by Anderssen and Jakeman (1975) the following procedure should be quite effective. If $z_i = g(t_i) + \epsilon_i$ is observed for $i = 1,2,\ldots,n$, then g is estimated by $g_{n,\lambda}$ and g' is estimated by $g'_{n,\lambda}$ as in Section 5.2.1. Since $g'_{n,\lambda}$ is a polynomial of degree 2 in each interval $[t_i,t_{i+1}]$, $g'_{n,\lambda}$ can be substituted into (5.2.4) and the integration carried out analytically. This is possible using formulas 129, 136, 153 of Pierce-Foster (1956). A similar method can be used in the Abel equations arising in computerized tomography, see Wahba (1980b).

In some important examples, g is a probability density function and only a random sample $X_1,\ldots,X_n$ from a population with density g is available. Some suggestions concerning this problem may be found in Wahba (1979d).

6. GCV in constrained regularization

6.1 Introduction

Westwater (1979) and Jackson (1979) have described experiments where outside information has greatly improved the estimate of the solution. For example, Westwater described how external measurement of the temperature inversion height aided in the determination of atmospheric temperature proviles. In geophysics, boundary values and moments may be known. Chambliss (1979), Merz (1979b), Wegman (1979) and Wright and Wegman (1980) have considered incorporating information that the true solution is non negative, monotone, or convex, in contexts similar to that discussed here. Sabatier (1977) considers positivity constraints from an entirely different point of view.

In this section, we consider some cases where f is known to satisfy certain side conditions. We discuss the insertion of these side conditions into the solution method, and extend the method of GCV to this case. The

simplest case is then f is known to satisfy ℓ linearly expressible side conditions, for example, boundary values, integrals, or moments of f may be known a priori. H must be chosen so that these side conditions are expressible as the values of continuous linear functionals $\{N_j\}_{j=1}^{\ell}$. If it is known that

$$N_j f = y_j,\ j = 1,2,\ldots,\ell,$$

exactly, then an estimate of f may be obtained by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}((Kf)(t_i)-z_i)^2 + \lambda||f||^2 \qquad (6.1.1)$$

subject to

$$N_j f = y_j,\ j = 1,2,\ldots,\ell. \qquad (6.1.2)$$

The set of $\{f\}$ in H satisfying (6.1.2) is closed and convex in H. More generally, we consider the case where f is known to be in a given closed, convex C in H. We consider the problem of minimizing (6.1.1) subject to $f\in C$. If H is a reproducing kernel space of functions defined on $[0,1]$, then the important case

$$C = \{f: f(t)\geq 0,\ t\in[0,1]\}$$

is included. The case

$$C = \{f:\ f^{(k)}(t)\geq\alpha(t),\ t\in[0,1]\}$$

is also included, if N_t defined by $N_t f = f^{(k)}(t)$ is a continuous linear functional in H for each $t\in[0,1]$. Thus, setting $k = 1$ or 2 we can require f to be monotone or convex.

We give a generalization of the GCV method for choosing λ in these constrained regulariation methods and discuss some computational strategies for fairly general C. We omit discussion of the case $||\cdot||$ is a semi norm, see Wahba (1979d) for some further details in that case.

6.2 GCV for constrained problems

It is easy to define "ordinary" cross validation for the problem: Find $f\epsilon H$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}((Kf)(t_i)-z_i)^2 + \lambda||f||^2 \qquad (6.2.1)$$

subject to

$$f\epsilon C,$$

where C is any closed convex set. Since (6.2.1) is a strongly convex functional on H for each fixed positive λ there always exists a unique minimizer $f_{n,\lambda}$ of (6.2.1) subject to $f\epsilon C$. Let $f_{n,\lambda}^{[k]}$ be the (unique) minimizer of

$$\frac{1}{n}\sum_{\substack{i=1\\ i\neq k}}^{n}((Kf)(t_i)-z_i)^2 + \lambda||f||^2 \qquad (6.2.2)$$

subject to $f\epsilon C$. Then the "ordinary" cross validation function, or "leaving out one" function $V_0(\lambda)$ may be defined analogously to (5.1.1) as

$$V_0(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Kf_{n,\lambda}^{[k]}(t_k)-z_k)^2 \qquad (6.2.3)$$

and the ordinary cross-validation estimate of λ is the minimizer of (6.2.3). However, $V_0(\lambda)$ will be prohibitive to compute in most cases, and it reduces in the unconstrained case to a procedure which can haye inferior properties to GCV. See GHW, CW.

Suppose that $Kf_{n,\lambda}(t_k)$ considered as a function of z_k is twice continuously differentiable in the neighborhood of z_k. Then, it is shown in the appendix that one can expand $Kf_{n,\lambda}(t_k)$ in a Taylor series about z_k and obtain

$$Kf_{n,\lambda}^{[k]}(t_k) = (Kf_{n,\lambda})(t_k) + \delta_k \frac{\partial}{\partial z_k}(Kf_{n,\lambda})(t_k) + O(\delta_k^2) \qquad (6.2.4)$$

where

$$\delta_k = (Kf_{n,\lambda}^{[k]})(t_k)-z_k$$

and $f_{n,\lambda}$ and $f_{n,\lambda}^{[k]}$ are the minimizers of (6.2.1) and (6.2.2) respectively, subject to $f\epsilon C$. Neglecting $O(\delta_k)$ in (6.2.4) results in the approximations

$$Kf_{n,\lambda}^{[k]}(t_k)-z_k \sim \frac{(Kf_{n,\lambda})(t_k)-z_k}{(1-a_{kk}(\lambda,z))}$$

and

$$V_0(\lambda) \approx \frac{1}{n}\sum_{k=1}^{n}\frac{((Kf_{n,\lambda})(t_k)-z_k)^2}{(1-a_{kk}(\lambda,z))^2} \qquad (6.2.5)$$

where

$$a_{kk}(\lambda,z) = \frac{\partial}{\partial z_k}(Kf_{n,\lambda})(t_k)\Big|_z.$$

(Compare (5.1.2) and (5.1.5), where $a_{kk}(\lambda,z)$ does not depend on z, since $Kf_{n,\lambda}(t_k)$ is a linear function of the data there, and the approximations are exact.) We believe that $V_0(\lambda)$ given by (6.2.3) should be replaced by the GCV function $V(\lambda)$ defined by

$$V(\lambda) = \frac{\frac{1}{n}\sum_{k=1}^{n}((Kf_{n,\lambda})(t_k)-z_k))^2}{(1-\frac{1}{n}\sum_{i=1}^{n}a_{kk}(\lambda,z))^2}. \qquad (6.2.6)$$

Provided that the map $A(\lambda)$: $E_n \to E_n$ which maps $z \to ((Kf_{n,\lambda})(t_1),\ldots,(Kf_{n,\lambda})(t_n))'$ is locally nearly linear and the δ_k are small, the same reasoning which led the substitution of $V_0(\lambda)$ by $V(\lambda)$ in the unconstrained case should work here. (See Wahba (1977), CW, GHW), and $V(\lambda)$ should provide an estimate of the λ which minimizes predictive mean square error. $V(\lambda)$ is also easier to compute, in general, than $V_0(\lambda)$.

6.3 The explicit form of the GCV function for a finite number of linear equality constraints

We suppose that H is an r.k.h.s. with r.k. $R(s,t)$, that N_j, $j = 1,2,\ldots,\ell$ are linearly independent continuous linear functionals on H, and it is known that $N_j f = y_j$, $j = 1,\ldots,\ell$, exactly. The representers $\{\psi_j\}$ of the (N_j) are those elements of H satisfying $N_j f = \langle \psi_j, f\rangle$, and by the properties of reproducing kernels, are given by

$$\psi_j(s) = N_{j(t)}R(s,t).$$

Here $N_{j(t)}$ means N_j applied to R considered as a function of t. The gram matrix M of the (ψ_j) is the $\ell \times \ell$ matrix with jkth entry

$$[M]_{jk} = \langle \psi_j, \psi_k\rangle = N_{j(s)}N_{k(t)}R(s,t)\ . \tag{6.3.1}$$

We assume that M is well conditioned, for practical purposes - otherwise some of the side conditions are redundant and should be eliminated. Then, the solution to the problem: Find $f \in H$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}((Kf)(t_i)-z_i)^2 + \lambda\|f\|^2$$

subject to

$$\langle \psi_j, f\rangle \equiv N_j f = y_j$$

is given by

$$f_{n,\lambda}(s) = \sum_{i=1}^{n} c_i \eta_i(s) + \sum_{j=1}^{\ell} b_j \psi_j(s) \tag{6.3.2}$$

where

$$\eta_i(s) = \int K(t_i,u)R(u,s)du$$

and c and b satisfy

$$\begin{pmatrix} K_n+n\lambda I & \vdots & L \\ \cdots & \cdots & \cdots \\ L' & \vdots & M \end{pmatrix}\begin{pmatrix} c \\ \cdots \\ b \end{pmatrix} = \begin{pmatrix} z \\ \cdots \\ y \end{pmatrix}$$

where the ijth entries of the matrices K_n and L are given by

$$[K_n]_{ij} = \langle \eta_i, \eta_j\rangle = \int\int K(t_i,u)R(u,v)K(t_j,v)dudv \tag{6.3.3}$$

$$[L]_{ij} = \langle \eta_i, \psi_j\rangle = N_{j(s)}\int K(t_i,u)R(s,u). \tag{6.3.4}$$

(See Kimeldorf and Wahba (1971).)

The solution $f_{n,\lambda}(s)$ is linear in the data z, and the $A(\lambda)$ operator (matrix) and hence the GCV function $V(\lambda)$ can be obtained by rewriting (6.3.2) in a form so that the dependence on z is more transparent. By using formulas for the inverse of a matrix in terms of submatrices, it can be shown that (6.3.2) is equivalent to

$$f_{n,\lambda}(s) = \sum_{i=1}^{n} \tilde{c}_i \xi_i(s) + \sum_{j=1}^{\ell} \tilde{b}_j \psi_j(s) \tag{6.3.5}$$

where

$$\begin{pmatrix} \xi_1(s) \\ \vdots \\ \xi_n(s) \end{pmatrix} = \begin{pmatrix} \eta_1(s) \\ \vdots \\ \eta_n(s) \end{pmatrix} - LM^{-1}\begin{pmatrix} \psi_1(s) \\ \vdots \\ \psi_\ell(s) \end{pmatrix}$$

and

$$M\tilde{b} = y \tag{6.3.6}$$
$$(\tilde{K}+n\lambda I)\tilde{c} = z - LM^{-1}y$$

with

$$\tilde{K} = K_n - LM^{-1}L'. \tag{6.3.7}$$

To obtain $V(\lambda)$ we then use the fact that

$$z - \begin{pmatrix} (Kf_{n,\lambda})(t_1) \\ \vdots \\ (Kf_{n,\lambda})(t_n) \end{pmatrix} = z - (LM^{-1}y) + \tilde{K}(\tilde{K}+n\lambda I)^{-1}(z-LM^{-1}y)$$

$$= (I-A(\lambda))(z-LM^{-1}y) \tag{6.3.8}$$

where

$$A(\lambda) = \tilde{K}(\tilde{K}+n\lambda I)^{-1} \quad . \tag{6.3.9}$$

Thus, $a_{kk}(\lambda,z) = \frac{\partial}{\partial z_k}(Kf_{n,\lambda})(t_k)$ is the kkth entry of $A(\lambda)$ of (6.3.9) and the GCV function of (6.2.6) is given by

$$V(\lambda) = \frac{\frac{1}{n}||(I-A(\lambda))(z-LM^{-1}y)||^2}{(\frac{1}{n}Tr(I-A(\lambda)))^2} \quad . \tag{6.3.10}$$

Letting $\tilde{K} = UDU'$ with $D = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$ and $x = (x_1,\ldots,x_n)' = U'(z-LM^{-1}y)$ gives

$$V(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}(\frac{n\lambda}{d_i+n\lambda})^2 x_i^2}{(\frac{1}{n}\sum_{i=1}^{n}\frac{n\lambda}{d_i+n\lambda})^2}$$

Before using this approach the experimenter should verify that errors in $LM^{-1}y$ are in fact entirely negligible compared to σ^2, and errors in computing M^{-1} are negligible. If this is the case, then the intrinsic rank of the problem is 2 + the intrinsic rank of D.

6.4 The GCV function for f in a closed convex set, computational strategies

We let $N_t, t\in[0,1]$ be a family of continuous linear functionals on $\mathcal{H}_R$, and suppose that f is known to satisfy

$$N_t f \geq \alpha(t),\ t\in[0,1], \tag{6.4.1}$$

where $\alpha(t)$ is a continuous function on $[0,1]$. We suppose that $N_t, t\in[0,1]$ is a "continuous with respect to t" family of continuous linear functionals, by this we mean that, if $(\psi_t, t\in[0,1])$ are the representers of $(N_t, t\in[0,1])$, $(\psi_t(s) = N_{t(v)}R(s,v))$ then

$$\lim_{\epsilon\to 0}||\psi_{t+\epsilon} - \psi_t|| = 0 \tag{6.4.2}$$

for each $t\in[0,1]$. (The important cases f non-negative, monotone, or convex can be included in this setup, by, for example, letting $\mathcal{H} = \mathcal{H}^m$ with $m = 1,2,$ or 3, respectively). With the assumption (6.4.2), it is intuitively reasonable that the set

$$C = \{f: N_t f \geq \alpha(t),\ t\in[0,1]\}$$

can be replaced for practical purposes by the set C_r,

$$C_r = \{f: N_{s_j} f \geq \alpha(s_j),\ j = 1,2,\ldots,r\},$$

where $s_1,\ldots,s_r$ is a sufficiently fine mesh in $[0,1]$. For related convergence theory, see Wahba (1973). Then, $f_{n,\lambda}$, the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}((Kf)(t_i)-z_i)^2 + \lambda||f||^2 \qquad (6.4.3)$$

subject to $f\in C_r$, is given by

$$f_{n,\lambda}(s) = \sum_{j=1}^{r} b_j \psi_{s_j}(t) + \sum_{i=1}^{n} c_i \eta_i(t) \qquad (6.4.4)$$

where b anc c are the solutions to the quadratic programming problem: Find b and c to minimize

$$\frac{1}{n}||Lb + K_n c - z||^2 + \lambda(b'Mb+2b'Lc+c'K_n c) \qquad (6.4.5)$$

subject to

$$Mb + L'c \geq \alpha \qquad (6.4.6)$$

where K_n, L and M are as in (6.3.1), (6.3.3) and (6.3.4) with $N_{j(s)}$ replaced by $N_{s_j(s)}$, etc., and $\alpha = (\alpha(s_1),\ldots,\alpha(s_r))'$ (see Kimeldorf and Wahba (1971). If n + r is under around 150 the solutions b and c to this problem can usually be obtained numerically, for fixed λ, from available library quadratic programming routines (for example, Madison Academic Computer Center (1977)).

The GCV function can be determined for given λ by observing that if $\ell \geq r$ constraints corresponding to a subset of $s_1,\ldots,s_r$ (call this subset $u_1,\ldots,u_\ell$), are <u>active</u> then the minimizer of (6.4.3) subject to $f\in C_r$ is the same as the minimizer of (6.4.3) subject to the <u>equality</u> constraints

$$\langle \psi_{u_j}, f\rangle = \alpha(u_j),\ j = 1,2,\ldots\ell. \qquad (6.4.7)$$

Thus, the locally linear dependence of $Kf_{n,\lambda}$ on z can be determined by looking at the dependence of $Kf_{n,\lambda}$ on z in the equality-constrained problem.

Considering the dependence of $(Kf_{n,\lambda})(t_k)$ on z, it is piecewise linear, the coefficients in the linear relationship changing only when a change in z causes a constraint to "just" become active or inactive. Thus (except for z in a measure 0 set of critical points) $V(\lambda)$ can be obtained as in Section 6.3 and is

$$V(\lambda) = \frac{\frac{1}{n}||(I-A_\ell(\lambda))(z-L_\ell M_\ell^{-1}\alpha_\ell)||^2}{(\frac{1}{n}Tr(I-A_\ell(\lambda)))^2}$$

where

$$A_\ell(\lambda) = \tilde{K}_\ell(\tilde{K}_\ell+n\lambda I)^{-1},$$

$$\alpha_\ell = (\alpha(u_1),\ldots,\alpha(u_\ell))$$

with

$$\tilde{K}_\ell = K_n - L_\ell M_\ell^{-1} L_\ell'$$

$$[L_\ell]_{ij} = \langle \eta_i, \psi_{u_j}\rangle \qquad \begin{matrix} i = 1,\ldots,n \\ j = 1,\ldots,\ell \end{matrix}$$

$$[M_\ell]_{ij} = \langle \psi_{u_i}, \psi_{u_j}\rangle \qquad \begin{matrix} i = 1,\ldots,\ell \\ j = 1,\ldots,\ell \end{matrix}$$

In an ill posed problem with noisy data the number of active constraints is likely to vary with λ, a small λ (little regularization) giving rise to a greater number of active constraints. It is possible that the solution to the unconstrained problem with the unconstrained GCV estimate of λ satisfies the constraints, while a smaller value of λ leads to active constraints and a smaller value of $V(\lambda)$. Enough trial values of λ must be taken so that the minimizer of $V(\lambda)$ can be obtained.

Various strategies suggest themselves for solving the constrained optimization problem iteratively by solving the unconstrained problem

and then sequentially picking up the most violated constraints until a solution in C_r is bound. An iterative procedure whereby it is only necessary to cary along two linear combinations of active constraints, has been developed by P.J. Laurent and the author, by using some results of Laurent and Martinet (1969), and will appear separately, see Laurent (1980), Wahba (1980a).

6.5 Other generalizations of the GCV method

The definition of the GCV function extends to other problems where $(Kf_{n,\lambda})(t_k)$ is not linear in the data.

There has been recent interest in robust smoothing, which is appropriate if the errors cannot be considered to have normal distributions, but may have outliers. For example, a robust smoothing spline is defined as the solution to: Find $f\in W_2^m$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n} \rho(f(t_i)-z_i) + \lambda\int_0^1 (f^{(m)}(u))^2 du \qquad (6.5.1)$$

where $\rho(\cdot)$ is a suitably chosen convex functional. One can define the GCV function exactly as in (6.2.6), or as

$$V(\lambda) = \frac{\frac{1}{n}\sum\rho(Kf_{n,\lambda})(t_i)-z_i)}{(1-\frac{1}{n}\sum_{i=1}^{n} a_{ii}(\lambda,z))^2}$$

where $a_{ii}(\lambda,z)$ is as in (6.2.5). See Huber (1979), Lenth (1979). When iterative methods are used to minimize (6.5.1) by iteratively replacing $(f(t_i)-z_i)$ by a quadratic local approximation, $A(\lambda,z)$ may be available at the last step of the iteration (O'Sulleabhain(1980)). For related numerical methods, see Utreras (1980). The definition (6.2.6) of $V(\lambda)$ is likely to be useful in some cases where K is a (mildly) nonlinear operator but we do not pursue this further here.

7. Acknowledgements

The author wishes to thank M.Z. Nashed for organizing the International Symposium on Ill Posed Problems, held at Newark, Delaware, October 2-6, 1979, where some of the ideas presented here had a chance to "ferment". She also wishes to acknowledge the help of conversations with Bob Anderssen and Don Chambless, and the support of the Office of Naval Research, without which much of this work would not exist.

APPENDIX

Theorem: Let $f_{n,\lambda}$ be the unique minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}[(Kf)(t_i)-z_i]^2 + \lambda||f||^2 \qquad (A.1)$$

subject to $f\in C$, and let $f_{n,\lambda}^{[k]}$ be the unique minimizer of

$$\frac{1}{n}\sum_{\substack{i=1\\ i\neq k}}^{n}[(Kf)(t_i)-z_i]^2 + \lambda||f||^2. \qquad (A.2)$$

Considering $(Kf_{n,\lambda})(t_k)$ as a function of $z = (z_1,\ldots,z_n)'$, suppose $(Kf_{n,\lambda})(t_k)$ is twice continuously differentiable with respect to z_k in the neighborhood of z. Then

$$(Kf_{n,\lambda}^{[k]})(t_k) = (Kf_{n,\lambda})(t_k) + \delta_k \frac{\partial}{\partial z_k}(Kf_{n,\lambda})\Big|_z + O(\delta_k^2), \qquad (A.3)$$

where

$$\delta_k = Kf_{n,\lambda}^{[k]}(t_k) - z_k.$$

Proof:

To indicate the dependence of $f_{n,\lambda}$ on z, write $f_{n,\lambda}(t,z) = f_{n,\lambda}(t)$. We first show

$$f_{n,\lambda}(t,z+\delta) \equiv f_{n,\lambda}^{[k]}(t),$$

where $\delta = (0,\ldots,0,\delta_k,0,\ldots,0)'$, δ_k is in the k^{th} position. To show this, denote $f_{n,\lambda}^{[k]}$ by h and $Kf_{n,\lambda}^{[k]}(t_k)$ by $\hat{z}_k$. Then

$$\frac{1}{n}\Big[\sum_{\substack{j=1\\ j\neq k}}^{n}(Kh)(t_j)-z_j)^2 + ((Kh)(t_k)-\hat{z}_k))^2\Big] + \lambda||h||^2$$

$$= \frac{1}{n}\sum_{\substack{j=1\\ j\neq k}}^{n}((Kh)(t_j)-z_j)^2 + \lambda||h||^2$$

$$< \frac{1}{n}\sum_{\substack{j=1\\ j\neq k}}^{n}((Kf)(t_j)-z_j)^2 + \lambda||f||^2, \text{ for any } f\in C \text{ other than } h$$

$$\leq \frac{1}{n}\Big[\sum_{\substack{j=1\\ j\neq k}}^{n}((Kf)(t_j)-z_j))^2 + ((Kf)(t_k)-\hat{z}_k)^2\Big] + \lambda||f||^2. \qquad (A.4)$$

Thus $h = f_{n,\lambda}^{[k]}$ is the minimizer of (A.4) which is also minimized uniquely by $f_{n,\lambda}(t,z+\delta)$.

It follows that

$$Kf_{n,\lambda}^{[k]}(t_k) \equiv (Kf_{n,\lambda})(t_k) + [(Kf_{n,\lambda})(t_k,z+\delta) - Kf_{n,\lambda}(t_k,z)]. \qquad (A.5)$$

Assuming that $(Kf_{n,\lambda})(t_k,z+\delta)$ is twice continuously differentiable in δ_k in the neighborhood of $\delta_k = 0$, expanding the second term on the right in (A.5) in a Taylor series in δ_k gives

$$Kf_{n,\lambda}^{[k]}(t_k) = (Kf_{n,\lambda})(t_k) + \delta_k \frac{\partial}{\partial z_k}(Kf_{n,\lambda})(t_k,z) + O(\delta_k^2).$$

REFERENCES

Akhiezer, N.I. and Glazman, I.M. (1961). Theory of Linear Operators in Hilbert Space, Volume I, Section 16. Translated from the Russian by M. Nestell. Ungar, New York.

Anderssen, R. S. (1976). Stable procedures for the inversion of Abel's equation. J. Inst. Maths. Applics. 17, 329-342.

Anderssen, R.S. and deHoog, F.R. (1979). Application and numerical solution of Abel-type integral equations. Talk at the International Symposium on Ill Posed Problems, Newark, Delaware, October 2-6, 1979, and manuscript.

Anderssen, R.S. and Jakeman, A.J. (1975). Abel type integral equations in stereology. II. Computational methods of solution and the random spheres approximation. J. Microscopy 105, pt. 2, 135-153.

Aronszajn, N. (1960). Theory of reproducing kernels. Transactions of The American Mathematical Society 68, 337-404.

Bjorck, A. and Elden, L. (1979). Methods in numerical algebra for ill posed problems. Talk at the International Conference Symposium on Ill Posed Problems. October 2-6, 1979. Newark, Delaware.

Chambless, D.A. (1979). Resolution Improvement of Gamma Radiation Spectometer Data. Talk at the International Symposium on Ill Posed Problems, Newark, Delaware, October 2-6, 1979.

Colli Franzone, P., Guerri, L., Taccardi, B. and Viganotti, C. (1979). The direct and inverse potential problems in electrocardiology. Numerical aspects of some regularization methods and application to data collected in isolated dog heart experiments. Laboratorio di Analisi Numerica del Consiglio Nazionale delle Richerche, Pavia, Italy Publicazioni N.222.

Craven, P. and Wahba (1979). "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross validation." Numer. Math. 31, 377-403.

Fleisher, J. (1979). Spline smoothing routines, reference manual for the 1110. Academic Computer Center, The University of Wisconsin-Madison.

Fritz, S., Wark, D.Q., Fleming, J.E., Smith, W.P., Jacobowitz, H., Hilleary, D.T. and Alishouse, J.C. (1972). Temperature sounding from satellites, NOAA Technical Report NESS 59, National Oceanic and Atmoshperic Administration, Washington, D.C.

Gehatia, M. and Wiff, D.R. (1970). "Solution of Fujita's equation for equilibrium sedimentation by applying Tikhonov's regularizing functions". J. Polymer Science A2, 9, 2039-2050.

Golub, G., Heath, M. and Wahba, G. (1979). "Generalized cross validation as a method for choosing a good ridge parameter", Technometrics 21, 2 215-223.

Huber, P. (1979). Robust smoothing, in "Robustness in Statistics", Proceedings of the Workshop at ARO, R. Launer and G. Wilkinson, eds. Academic Press.

IMSL (1980). International Mathematical and Statistical Libraries, Inc., Edition 8, Houston, TX, subroutine ICSSCV

Jackson, D.D. (1979). Use of prior data in geophysical inverse problems. Talk at the International Symposium on Ill Posed Problems, Newark, Delaware, October 2-6, 1979.

Jakeman, A.J. and Anderssen, R.S. (1975a). "Abel type integral equations in stereology". I. General discussion. J. Microscopy, 106, pt. 2, 121-133.

Jakeman, R.J. and Anderssen, R.S. (1975b). On optimal forms for stereological data, in Proceedings of the Fourth International Congress for Stereology, Gaithersburg, MD, September 1975.

Kimeldorf, G. and Wahba, G. (1971). "Some results on Tchebycheffian spline functions. J. Math. Anal. Applic. 31, 1, 82-95.

Laurent, P.J. (1980). An algorithm for the computation of spline functions with inequality constraints (unpublished manuscript).

Laurent, P.J. and Martinet, B. (1969). Methodes duales pour le calcul du minimum d'une fonction convexe sur une intersection de convexes. Proceedings of the Symposium on Optimization, held in Nice, France, June 23-July 5, 1969, Lecture Notes in Mathematics #132, pp. 159-180, (1970) Springer-Verlag.

Lenth, R. (1979). Robust Smoothing Splines. Talk at the Institute of Mathematical Statistics Special Topics Meeting on Splines and Approximations, Boulder, CO, October 25-26, 1979.

Madison Academic Computing Center (1977). QUADPR/QUADMP Quadratic Programming Subroutines, University of Wisconsin-Madison MACC Report 1506.

Merz, P.H. (1978). Spline smoothing by generalized cross-validation, a technique for data smoothing. Chevron Research Corp. Richmond, CA.

Merz, P.H. (1979a). Determination of adsorption energy distribution by regularization and a characterization of certain adsorption isotherms. Chevron Research Corp., Richmond, CA.

Merz, P.H. (1979b). Talk at the International Symposium on Ill Posed Problems, Newark, Delaware, October 2-6, 1979.

Nashed, M.Z. (1979a). Projection methods for ill posed integral and operator equations. Talk at the International Symposium on Ill Posed Problems, Newark, Delaware, October 2-6, 1979.

Nashed, M.Z. (1979b). Ill posed problems in system analysis and identification, in Proceedings of the Symposium on Information and System Sciences. Johns Hopkins University, 1979.

O'Sulleabhain, S.F. (1980). Personal communication.

Pierce, B.O. and Foster, R.M. (1956). A Short Table of Integrals, Ginn and Co., Boston.

Reinsch, C.M. (1967). "Smoothing by spline functions", Numer. Math. 10, 177-183.

Sabatier, P.C. (1977). Positivity constraints in linear inverse problems-II. Applications. Geophys. J.R. Astr. Soc. 48, 443-469.

Smith, W.L. and Woolf, H.M. (1976). "The use of eigenvectors of statistical covariance matrices for interpreting satellite sounding radiometer observations. J. Atmospheric Sciences 33, 7, 1127-1140.

Stutzle, W. (1977). Estimation and parameterization of growth curves. Doctoral disseration, ETH, Zurich.

Utreras, F. (1979a). Utilization de la methode de validation croisee pour le lissage par fonctions spline a une ou deux variables. Universite Scientifique et Medicale de Grenoble, Thesis.

Utreras, F. (1979b). Cross validation techniques for smoothing spline functions in one or two dimensions. In "Smoothing Techniques for Curve Estimation", T. Gasser and M. Rosenblatt, eds. Lecture Notes in Mathematics, No. 757, Springer-Verlag, Berlin.

Utreras, F. (1980). On computing robust splines and applications. Manuscript, Departamento de mathematicas, Universidad de Chile, Santiago.

Wahba, G. (1973). "On the minimization of a quadratic functional subject to a continuous family of linear inequality constraints", SIAM J. Control, 11, 1, 64-79.

Wahba, G. (1977). "Practical approximate solutions to linear operator equations when the data are noisy", SIAM J. Numerical Analysis, 14, 4.

Wahba, G. (1979a). "How to smooth curves and surfaces with splines and cross-validation. In Proceedings of the 24th Conference on the Design of Experiments: U.S. Army Research Office, Report 79-2, also University of Wisconsin-Madison, Statistics Department Technical Report No. 555.

Wahba, G. (1979b). Convergence rates of "Thin Plate" smoothing splines when the data are noisy, in "Smoothing Techniques for Curve Estimation", T. Gasser and M. Rosenblatt, eds. Lectures notes in Mathematics, No. 757, Springer-Verlag, Verlin.

Wahba, G. (1979c). Smoothing and ill posed problems, in "Solution Methods for Integral Equations with Applications: Michael Golberg, ed. Plenum Press.

Wahba, G. (1979d). Ill posed problems: Numerical and statistical methods for mildly, moderately, and severely ill posed problems with noisy data. University of Wisconsin-Madison, Statistics Department Technical Report No. 595.

Wahba, G. (1980a). Positively/convexly constrained smoothing problems. Abstract SIAM 1980 Fall meeting, Houston, Texas.

Wahba, G. (1980b). A new approach to the numerical evaluation of the inverse radon transform with discrete, noisy data. University of Wisconsin-Madison Statistics Department Technical Report No. 612, to appear in "Mathematical Aspects of Computerized Tomography", F. Natterer, Ed., Springer-Verlag, Berlin.

Wahba, G. and Wendelberger, J. (1980). "Some new mathematical methods for variational objective analysis using splines and cross validation". Monthly Weather Review, 108, 8, 1122-1143.

Wegman, E. (1979). Two approaches to non parametric regression: splines and isotonic inference. Talk at the Japan Statistical Society Meeting in Tokyo, November 28-December 1, 1979, also manuscript.

Wendelberger, J. (1980). University of Wisconsin-Madison, Statistics Department, Thesis, to appear.

Welch, J. (1979). Personal communication.

Westwater, E.R. (1979). Ill posed problems in remote sensing of the earth's atmoshpere by microwave radiometry. Talk at the International Conference on Ill Posed Problems, Newark, Delaware, October 2-6, 1979.

Wright, I. and Wegman, E. (1980). Isotonic, convex and related splines. Ann. Statist. 8, 5, 1023-1035.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered) UWIS-DS-80-629

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report No. 629
2. GOVT ACCESSION NO.: AD-A097027
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): CROSS VALIDATION AND CONSTRAINED REGULARIZATION METHODS FOR MILDLY ILL POSED PROBLEMS, REVISION
5. TYPE OF REPORT & PERIOD COVERED:
6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Grace Wahba
8. CONTRACT OR GRANT NUMBER(s): N00014-77-C-0675
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, University of Wisconsin, Madison, WI 53706
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Arlington, VA
12. REPORT DATE: 1 Jan 1980
13. NUMBER OF PAGES: 36
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): 12 22
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:
16. DISTRIBUTION STATEMENT (of this Report): Distribution of this document is unlimited
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):
18. SUPPLEMENTARY NOTES: None
19. KEY WORDS (Continue on reverse side if necessary and identify by block number): ill posed problems; constrained regularization; cross validation
20. ABSTRACT (Continue on reverse side if necessary and identify by block number): (see reverse side)

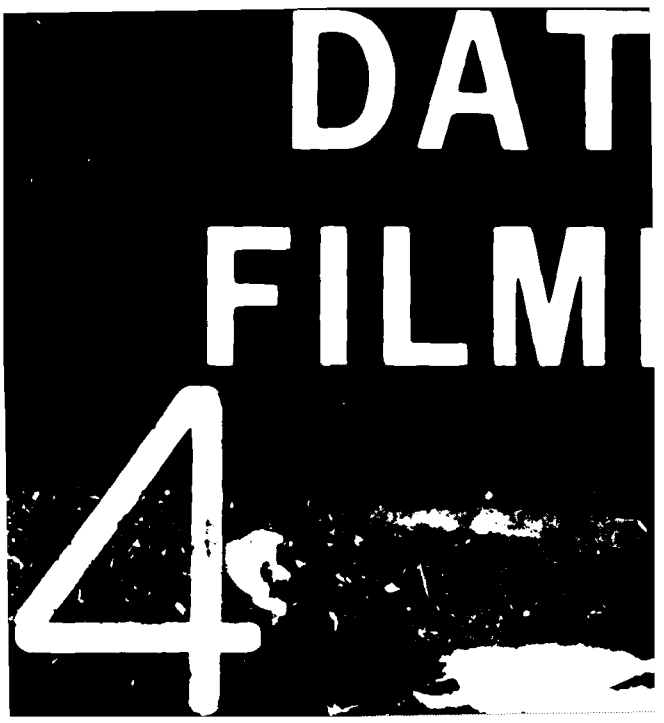
DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD A097027

ABSTRACT

The model $z_i = \int_0^1 K(t_i,s)f(s)ds + \epsilon_i$, $i = 1,2,\ldots,n$, is studied, where K is assumed known, the (ϵ_i) are random errors, and it is desired to estimate f from the data (z_i). The (regularized) estimate of f will be taken as the minimizer in H of $\frac{1}{n}\sum_{i=1}^{n}((Kf)(t_i)-z_i)^2 + \lambda||f||^2$ (=(*)), where H is a reproducing kernel space with norm $||\cdot||$. We first define the intrinsic rank of this experiment. This definition is used to provide insight into the circumstances in which one may expect to estimate f well, moderately well, or poorly. The sensitivity of a regularized estimate to the (ϵ_i) is made explicit. After giving the intrinsic rank of the examples of first and second derivative, Abel's equation and Fujita's equation, it is argued that the first three are only mildly ill posed and if f is "nice" it will be amenable to accurate estimation. The method of generalized cross validation for choosing λ is reviewed and numerical results for the estimation of first and second derivative from noisy data are given. A method for solving Abel's equations when inversion formulae are available is described which exploits properties of the above mentioned first derivative estimate. The use of outside information in the estimation of f is then considered. We consider the cases where f is known to satisfy a finite number of linear equality constraints (for example, boundary values, or moment conditions) or a continuous family of linear inequality constraints (including positivity, monotonicity or convexity). f is then estimated by minimizing or approximately minimizing (*), subject to these constraints. We extend the generalized cross validation method for choosing λ to this case. Finally, the extension of this method to certain robust estimation problems is suggested.


DAT
FILM
4